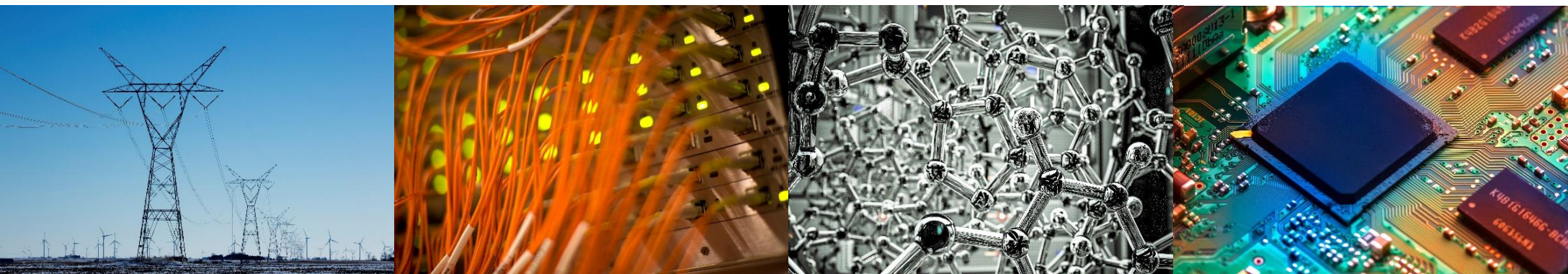


# Multilingual Cross-lingual and Cross-tonal Transfer for Under-resourced Asian Languages on CTC-based Acoustic Modeling

Jialu Li



**I** ILLINOIS

Electrical & Computer Engineering

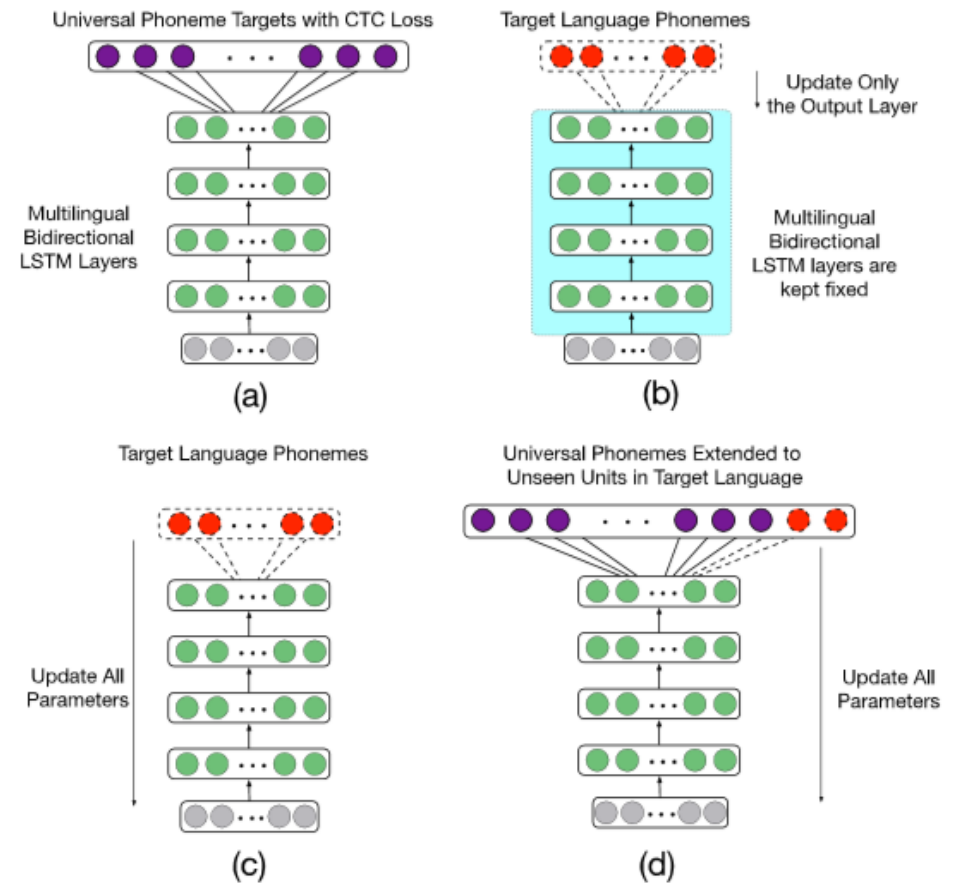
GRAINGER COLLEGE OF ENGINEERING

# Outline

- Introduction
- Data
- Model
  - IPA-based phones and tones
- Experiments
- Results
- Discussions
- Future work

# Introduction

- Automatic speech recognition (ASR) for under-resourced languages has been a challenging area
- Universal phone-based multilingual training and adapting is an effective method for neural network-based model
- Some tone recognition experiments have done in the past, but few experiments have addressed the cross-tonal effects in ASR study



Example of universal phone-based multilingual adaptation scheme

# Data

- Training languages
  - Mandarin from 1997 Mandarin Broadcast News Speech corpus
  - Cantonese, Vietnamese, Javanese, Swahili and Turkish from BABEL speech corpus
- Testing languages
  - Lao and Zulu from BABEL speech corpus
  - English from TED-LIUM dataset

# Data

- Audio time distributions for training languages in hours
  - Each of dev and test sets are composed of 800 utterances

| Languages  | Train | Dev  | Test |
|------------|-------|------|------|
| Mandarin   | 26.50 | 1.46 | 1.43 |
| Cantonese  | 31.08 | 1.56 | 1.84 |
| Vietnamese | 18.24 | 1.54 | 1.39 |
| Javanese   | 8.91  | 3.15 | 1.75 |
| Swahili    | 9.22  | 1.82 | 1.81 |
| Turkish    | 27.75 | 1.23 | 1.80 |

- To simulate under-resourced condition, audio data is limited to 1 hour for training, development and testing set for all testing languages

# Data

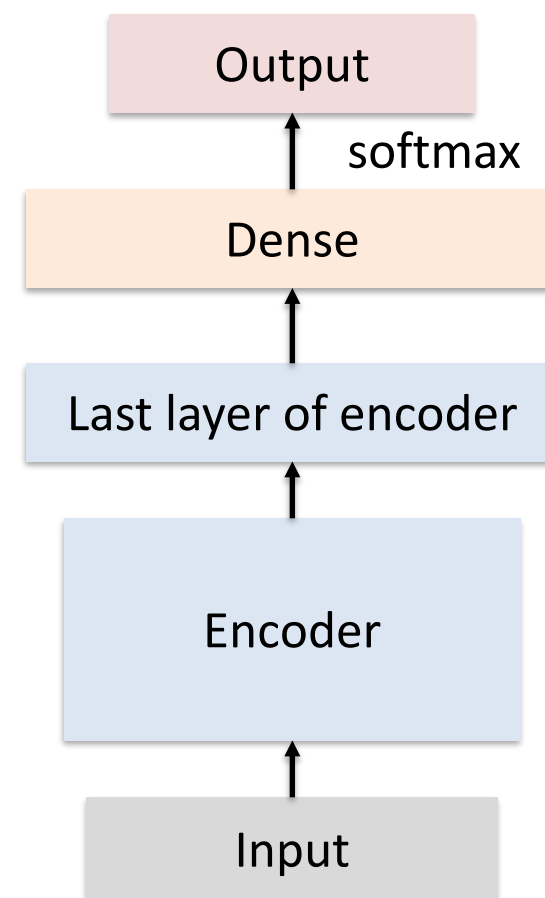
- Preprocessing of audio data
  - 40-dimensional log Mel filterbank features are used
  - 25 ms Hamming window with 10 ms shift over each audio segment
  - Z-normalization is performed over each speaker
  - 40-dimensional log Mel filterbank features + F0 are currently training
- Preparation of phonetic transcriptions
  - LanguageNet Grapheme-to-Phoneme (G2P) transducers are used to convert provided grapheme transcripts to phoneme transcripts
  - Phoneme transcripts are International phonetic alphabets (IPA)-based phones

# Model

- CTC-based Acoustic Modeling
  - Encoder learns acoustic features
  - Maximum likelihood sequences over outputs
  - Loss objectives

$$\mathcal{L}_{mle} = - \sum_{y \in |S|} \log(p(\mathbf{y}|\mathbf{X})) = - \sum_{y \in |S|} \log \prod_{t=1}^n p(y_t|\mathbf{x}_t)$$

- Outputs collapsed blank symbol and duplicated characters
  - aa-- → a
  - a-aa → aa



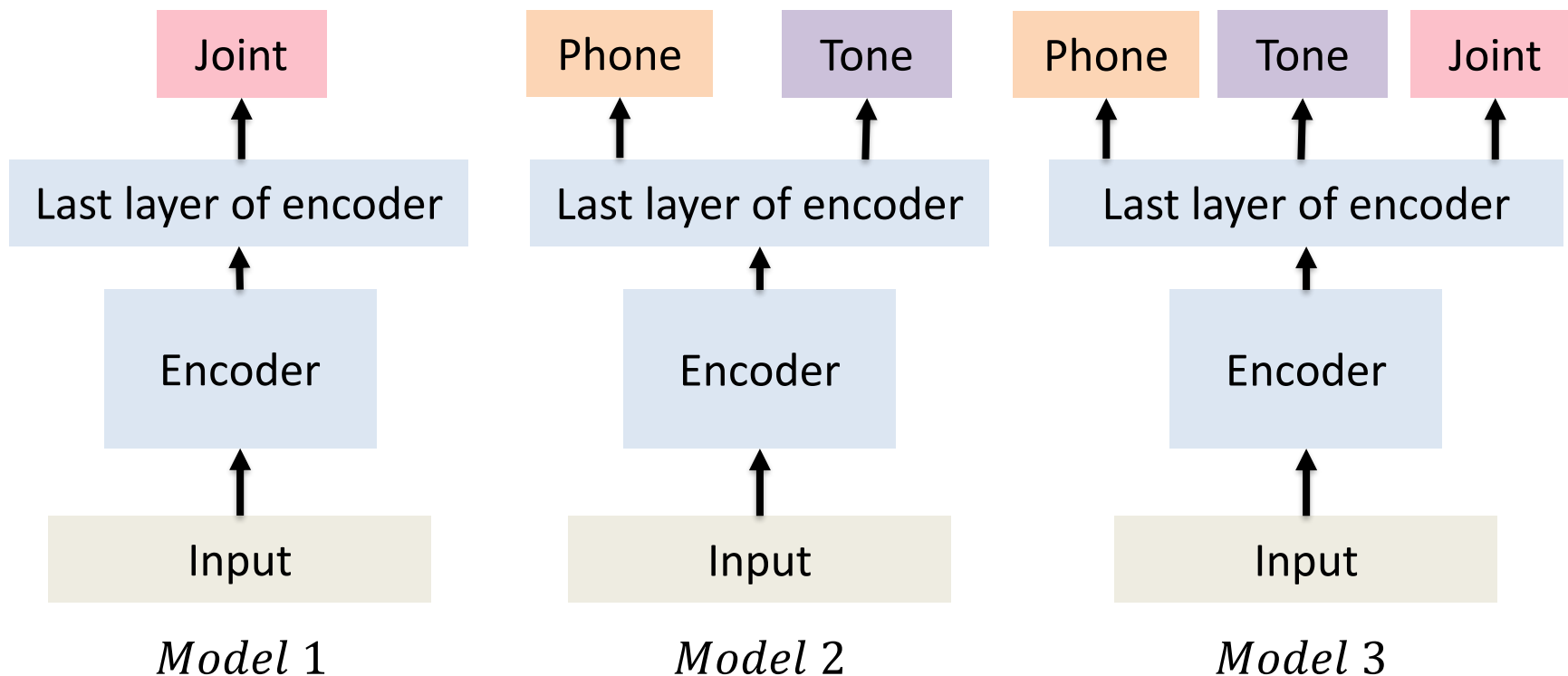
# Model

- Multi-task learning

$$\mathcal{L}_1 = \mathcal{L}_{ctc\_joint}$$

$$\mathcal{L}_2 = \mathcal{L}_{ctc\_phone} + \mathcal{L}_{ctc\_tone}$$

$$\mathcal{L}_3 = \mathcal{L}_{ctc\_phone} + \mathcal{L}_{ctc\_tone} + \mathcal{L}_{ctc\_joint}$$



# IPA-based tones

## Level tones

|       |   |                    |
|-------|---|--------------------|
| ◌̊    | ┐ | Extra-high (top)   |
| ◌̊̄   | ┐ | High               |
| ◌̄    | ┐ | Mid                |
| ◌̌    | ┐ | Low                |
| ◌̋    | ┘ | Extra-low (bottom) |
| ↓ ◌̊̄ |   | Downstep           |
| ↑ ◌̊̄ |   | Upstep             |

## Tones and Accents

### Level

### Contour

|        |            |        |                |
|--------|------------|--------|----------------|
| ě or ┐ | Extra high | ě or ┘ | Rising         |
| é      | High       | ê      | Falling        |
| ē      | Mid        | ┐┐     | High rising    |
| è      | Low        | ┘┘     | Low rising     |
| ě      | Extra low  | ┐┐┐    | Rising-falling |
| ↓      | Downstep   | ↗      | Global rise    |
| ↑      | Upstep     | ↘      | Global fall    |

[https://en.wikipedia.org/wiki/International\\_Phonetic\\_Alphabet\\_chart](https://en.wikipedia.org/wiki/International_Phonetic_Alphabet_chart)

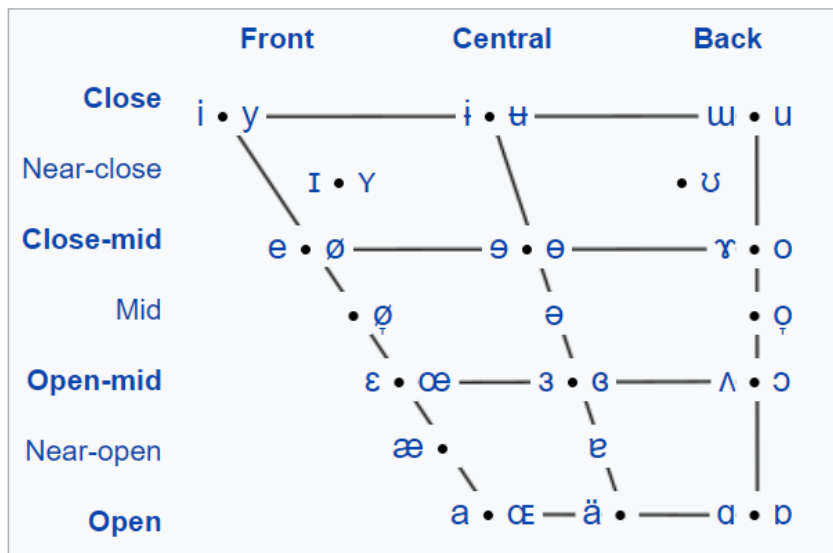
# Model - adapting IPA-based tones

- Example of IPA-based tone letters used in training and testing languages

| Man     | Descriptions       | Can   | Descriptions | Viet    | Descriptions       | Lao   | Descriptions |
|---------|--------------------|-------|--------------|---------|--------------------|-------|--------------|
| a 1     | High level         | a 1   | High level   | a 1     | Ngang; mid level   | a 1   | Low level    |
| a 1 1   | mid-rising         | a 1   | Mid level    | a 1 1   | Sac; mid-rising    | a 1   | Mid level    |
| a 1 1 1 | Low-dipping rising | a 1   | Low level    | a 1 1 1 | Nang; mid-falling  | a 1   | High level   |
| a 1 1   | Falling            | a 1 1 | Mid-rising   | a 1 1 h | hoi; mid-falling   | a 1 1 | Low-rising   |
|         |                    | a 1 1 | Low-falling  | a 1 1 1 | Huyen; low-falling | a 1 1 | High-falling |
|         |                    | a 1 1 | Low-rising   | a 1 1 1 | Nga; mid-rising    | a 1 1 | Mid-falling  |
|         |                    |       |              |         |                    | ?     | Stop glottal |
|         |                    |       |              |         |                    | h     | Stop glottal |

# IPA-based phones

## Vowels [\[ edit \]](#)



Vowels beside dots are:  
unrounded • rounded

[IPA help](#) • [audio](#) • [full chart](#) • [template](#)

## Pulmonic consonants [\[ edit \]](#)

| Manner ↓               | Labial                  |                         | Coronal               |                         |                         |                         | Dorsal                  |                         |                         | Laryngeal               |                         |
|------------------------|-------------------------|-------------------------|-----------------------|-------------------------|-------------------------|-------------------------|-------------------------|-------------------------|-------------------------|-------------------------|-------------------------|
|                        | Bilabial                | Labio-dental            | Linguo-labial         | Dental                  | Alveolar                | Post-alveolar           | Retro-flex              | Palatal                 | Velar                   | Uvular                  | Pharyngeal/epiglottal   |
| Nasal                  | $\text{m}$ $\text{m}$   | $\text{ɱ}$              | $\text{ɱ}$            |                         | $\text{n}$ $\text{n}$   |                         | $\text{ɳ}$ $\text{ɳ}$   | $\text{ɲ}$ $\text{ɲ}$   | $\text{ŋ}$ $\text{ŋ}$   | $\text{ɴ}$              |                         |
| Stop                   | $\text{p}$ $\text{b}$   | $\text{ɸ}$ $\text{β}$   | $\text{ɸ}$ $\text{β}$ | $\text{t}$ $\text{d}$   | $\text{t}$ $\text{d}$   | $\text{tʃ}$ $\text{dʒ}$ | $\text{ʈ}$ $\text{ɖ}$   | $\text{c}$ $\text{ɟ}$   | $\text{k}$ $\text{g}$   | $\text{q}$ $\text{ɢ}$   | $\text{ʕ}$ $\text{ʕ}$   |
| Sibilant affricate     |                         |                         |                       |                         | $\text{ts}$ $\text{dz}$ | $\text{tʃ}$ $\text{dʒ}$ | $\text{ʈʂ}$ $\text{ɖʒ}$ | $\text{tɕ}$ $\text{dʑ}$ |                         |                         |                         |
| Non-sibilant affricate | $\text{pɸ}$ $\text{bβ}$ | $\text{ɸɸ}$ $\text{ββ}$ |                       | $\text{tθ}$ $\text{dð}$ | $\text{tʃ}$ $\text{dʒ}$ | $\text{tʃ}$ $\text{dʒ}$ |                         | $\text{cç}$ $\text{jʝ}$ | $\text{kx}$ $\text{gɣ}$ | $\text{qχ}$ $\text{ɢʁ}$ | $\text{ʕʕ}$ $\text{ʕh}$ |
| Sibilant fricative     |                         |                         |                       |                         | $\text{s}$ $\text{z}$   | $\text{ʃ}$ $\text{ʒ}$   | $\text{ʂ}$ $\text{ʐ}$   | $\text{ç}$ $\text{ʝ}$   |                         |                         |                         |
| Non-sibilant fricative | $\text{ɸ}$ $\text{β}$   | $\text{f}$ $\text{v}$   | $\text{ɸ}$ $\text{β}$ | $\text{θ}$ $\text{ð}$   | $\text{tʃ}$ $\text{dʒ}$ | $\text{tʃ}$ $\text{dʒ}$ | $\text{ʈ}$ $\text{ɖ}$   | $\text{ç}$ $\text{ʝ}$   | $\text{x}$ $\text{ɣ}$   | $\text{χ}$ $\text{ʁ}$   | $\text{ħ}$ $\text{ʕ}$   |
| Approximant            |                         | $\text{ʋ}$ $\text{ʋ}$   |                       |                         | $\text{j}$ $\text{j}$   |                         | $\text{ɻ}$ $\text{ɻ}$   | $\text{j}$ $\text{j}$   | $\text{ɰ}$ $\text{ɰ}$   |                         |                         |
| Tap/flap               |                         | $\text{v}$ $\text{v}$   | $\text{ɾ}$            | $\text{ɾ}$ $\text{r}$   | $\text{ɾ}$ $\text{r}$   | $\text{ɾ}$ $\text{r}$   |                         |                         |                         | $\text{ɢ}$              | $\text{ʕ}$              |
| Trill                  | $\text{ʙ}$ $\text{ʙ}$   |                         |                       |                         | $\text{r}$ $\text{r}$   | $\text{ɽ}$ $\text{ɽ}$   |                         |                         |                         | $\text{ʀ}$ $\text{ʀ}$   | $\text{ħ}$ $\text{ʕ}$   |
| Lateral affricate      |                         |                         |                       |                         | $\text{tɬ}$ $\text{dɮ}$ | $\text{tɬ}$ $\text{dɮ}$ |                         | $\text{cɭ}$ $\text{jɥ}$ | $\text{kɭ}$ $\text{gɮ}$ |                         |                         |
| Lateral fricative      |                         |                         |                       |                         | $\text{ɬ}$ $\text{ɮ}$   | $\text{ɬ}$ $\text{ɮ}$   | $\text{ɬ}$ $\text{ɮ}$   | $\text{ɬ}$ $\text{ɮ}$   | $\text{ɬ}$ $\text{ɮ}$   |                         |                         |
| Lateral approximant    |                         |                         |                       |                         | $\text{l}$ $\text{l}$   | $\text{l}$ $\text{l}$   | $\text{l}$ $\text{l}$   | $\text{l}$ $\text{l}$   | $\text{l}$ $\text{l}$   | $\text{l}$ $\text{l}$   |                         |
| Lateral tap/flap       |                         |                         |                       |                         | $\text{l}$              | $\text{l}$              | $\text{l}$              | $\text{l}$              | $\text{l}$              | $\text{l}$              |                         |

[https://en.wikipedia.org/wiki/International\\_Phonetic\\_Alphabet\\_chart](https://en.wikipedia.org/wiki/International_Phonetic_Alphabet_chart)

# Model- adapting IPA-based phones

- Denote the fully connected weight matrix as  $W \in |S| \times d$ , where  $|S|$  is the size of output label set and  $d$  is the hidden dimension of weight matrix.

$$W_k = \begin{cases} \frac{\sum_{j=1}^n W_{jk}}{\sum_{j=1}^n \mathbb{1}_{k \in S_j}} & \text{if } \sum_{j=1}^n \mathbb{1}_{k \in S_j} \neq 0 \\ \frac{\sum_{j=1}^n W_{jk'}}{\sum_{j=1}^n \mathbb{1}_{k' \in S_j}} & \text{else} \end{cases}$$

- $k$ : adapted phone
  - $k'$ : the closest phone of  $k$
  - $|S_j|$ : output label set for training language  $j$
  - $n$ : number of training languages
- Example of closest phone pairs (a, a:) and (ʌ, ɔ)

## Suprasegmentals

|    |                              |
|----|------------------------------|
| '  | Primary stress               |
| ˈ  | Secondary stress             |
| ː  | Long                         |
| ˑ  | Half-long                    |
| ◌̥ | Extra-short                  |
|    | Minor (foot) group           |
|    | Major (intonation) group     |
| .  | Syllable break               |
| ◌̣ | Linking (absence of a break) |

# Model

- Sample output label for three multi-task learning tiers
  - For a single vowel, a ɪ ɪ
    - Phone tier: single phone without tone letters (a)
    - Tone tier: tones assigned to this vowel (ɪ ɪ)
    - Joint tier: complex units combined phone and tone letters (a1)
  - For a consonant, b
    - Phone tier: single phone without tone letters (b)
    - Tone tier: consonants are treated as word boundary (<sym\_boundary>)
    - Joint tier: same as the phone tier (b)

# Experiments

- **Experimental setup**

- CTC model was implemented on the eXtensible Neural Machine Translation Toolkit (XNMT) via the framework Dynet
- Encoder: 3 layers of bi-directional pyramidal LSTM with hidden dimension 512
- Optimizer: Adadelta with learning rate 0.004
- Early stopping criteria was used to select the best model
- Beam width: 25

- **Baseline adapted model**

- Kept the original CTC model and trained it on 1 hour limited data
- Reduced hidden dimension from 512 to 2 or 4

# Results

## ■ Output error rates

Table 3: *Phone error rates (PER), tone error rates (TER), and joint phone and tone error rates (JER) for multilingual training models*

|      | Model1<br>JER | Model2<br>PER TER |        | Model3<br>PER TER JER |        |        |
|------|---------------|-------------------|--------|-----------------------|--------|--------|
| Man  | 52.22%        | 54.73%            | 58.06% | 47.46%                | 61.14% | 56.35% |
| Can  | 36.55%        | 41.69%            | 45.19% | 34.90%                | 40.91% | 36.73% |
| Viet | -             | 51.57%            | 48.47% | 45.74%                | 43.50% | 44.54% |
| Jav  | 51.26%        | -                 | -      | -                     | -      | -      |
| Swa  | 43.20%        | -                 | -      | -                     | -      | -      |
| Tur  | 58.19%        | -                 | -      | -                     | -      | -      |

Table 4: *Phone error rates (PER), tone error rates (TER), and joint error rates (JER) for adapted and baseline models for testing languages.*

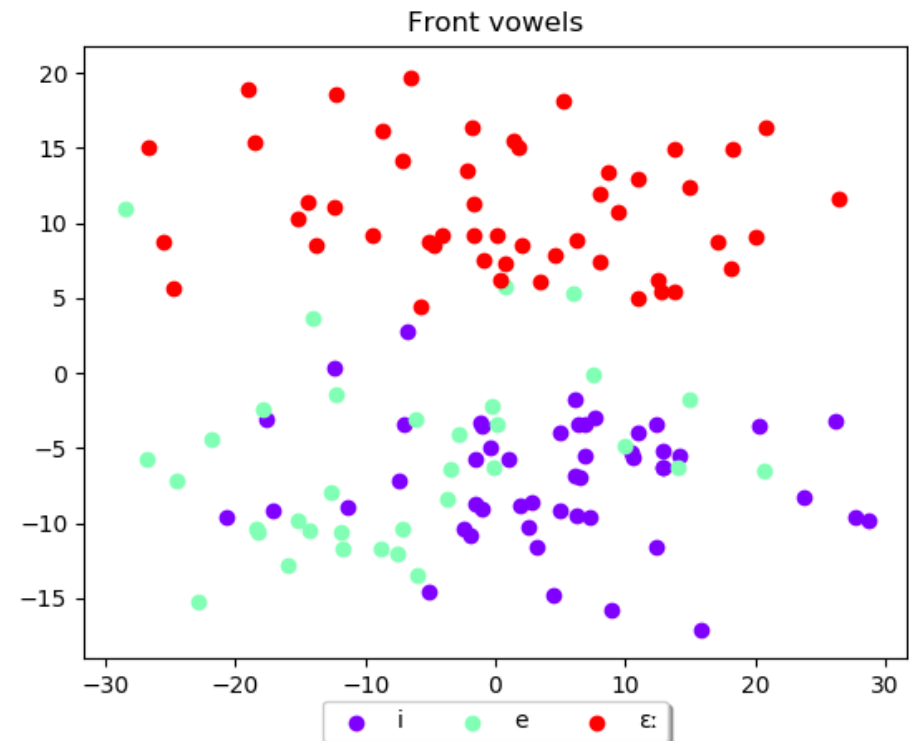
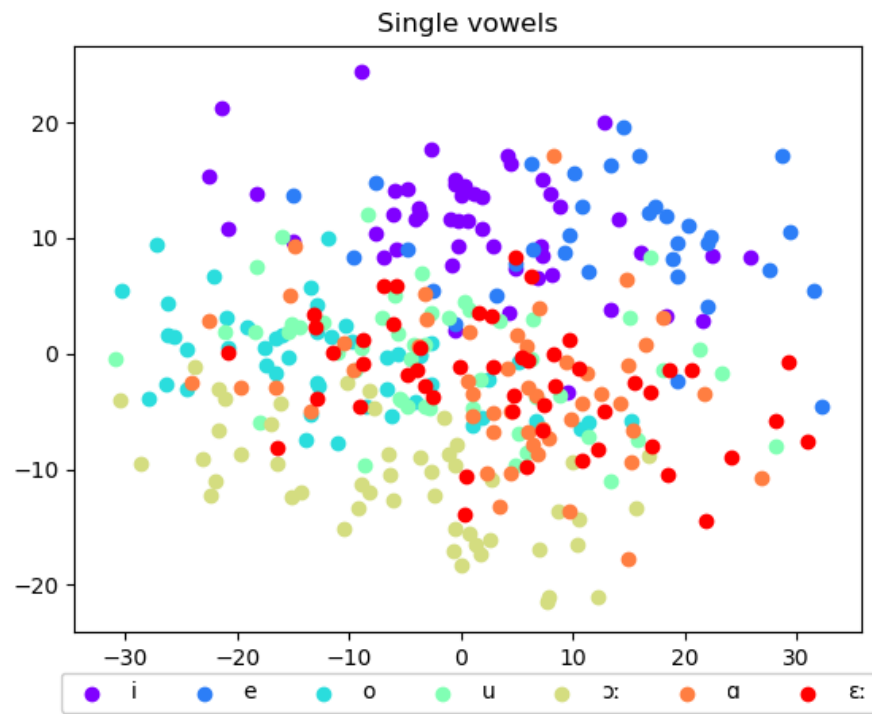
|         | Model1<br>JER | Model2<br>PER TER |        | Model3<br>PER TER JER |        |        |
|---------|---------------|-------------------|--------|-----------------------|--------|--------|
| Lao     | 48.79%        | 51.23%            | 44.38% | 51.25%                | 43.71% | 50.19% |
| Lao-bl  | 83.81%        | 90.05%            | 79.01% | 90.53%                | 82.52% | 82.36% |
| Zulu    | 66.56%        | -                 | -      | -                     | -      | -      |
| English | 77.08%        | -                 | -      | -                     | -      | -      |

# Discussions

- Visualizations of hidden vectors from CTC output weight matrixes based on best path search results for all three multi-task learning tiers in Model3
- Principle component analysis (PCA) are used to reduce hidden dimension from 512 to 50, then t-distributed Stochastic Neighbor Embedding (t-SNE) is used to cluster reduced vectors

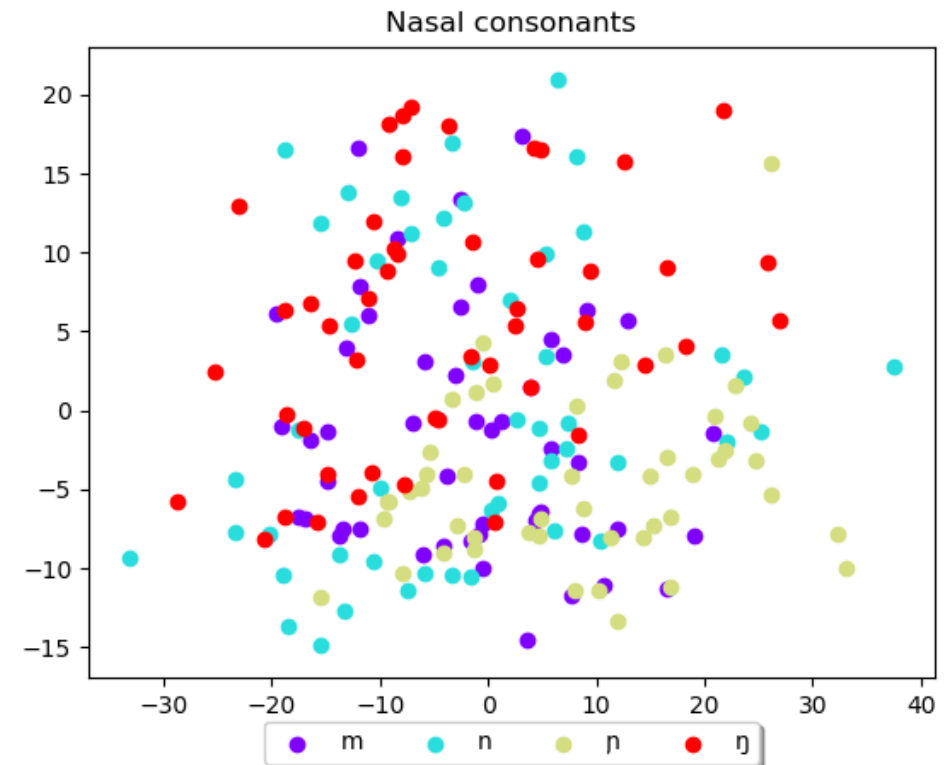
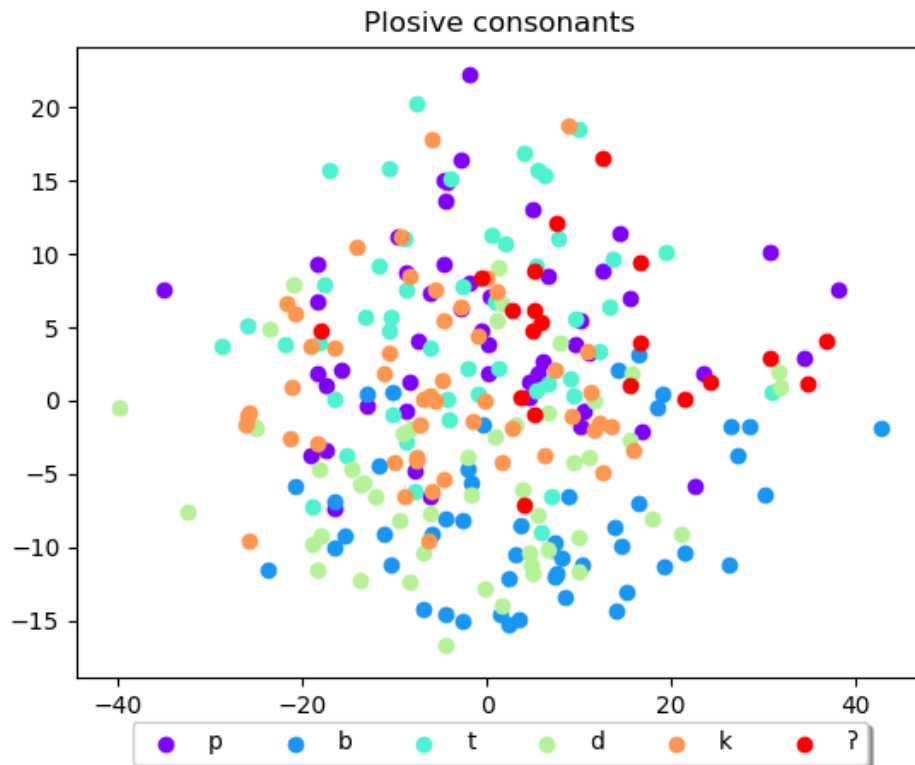
# Discussions

- Phone tier: vowels



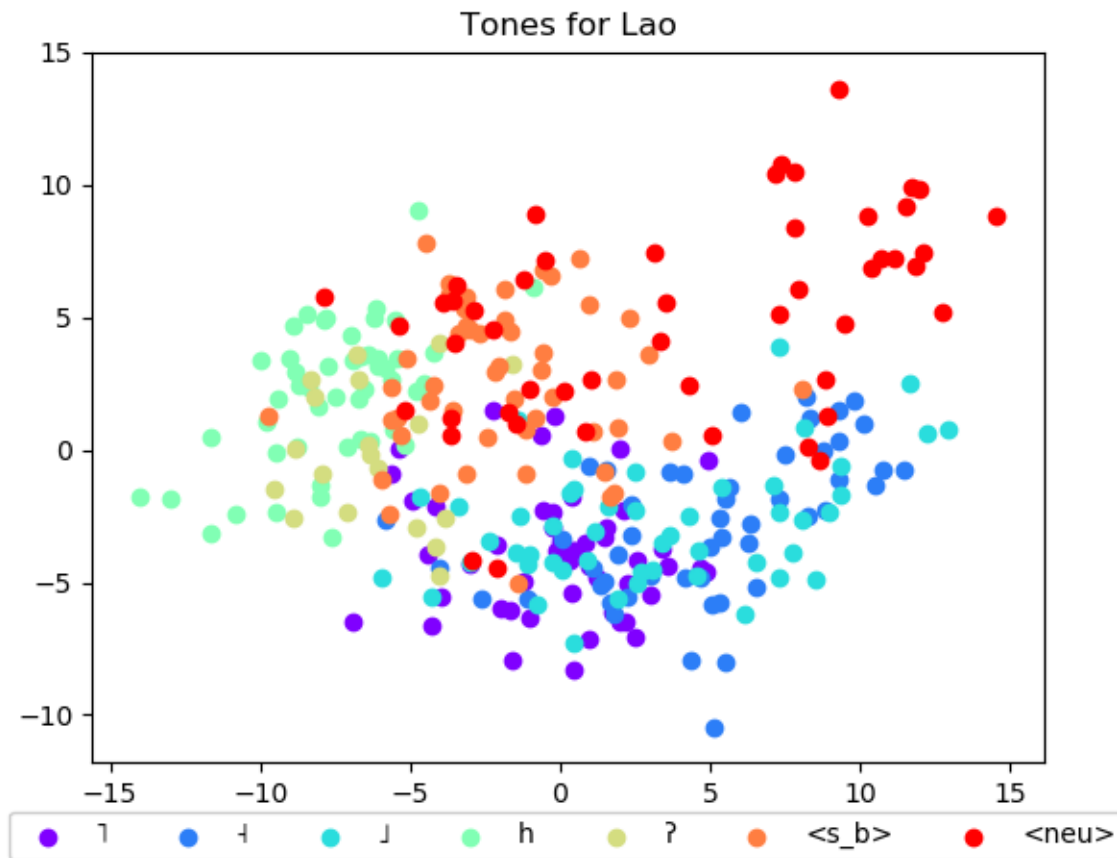
# Discussions

- Phone tier: consonants



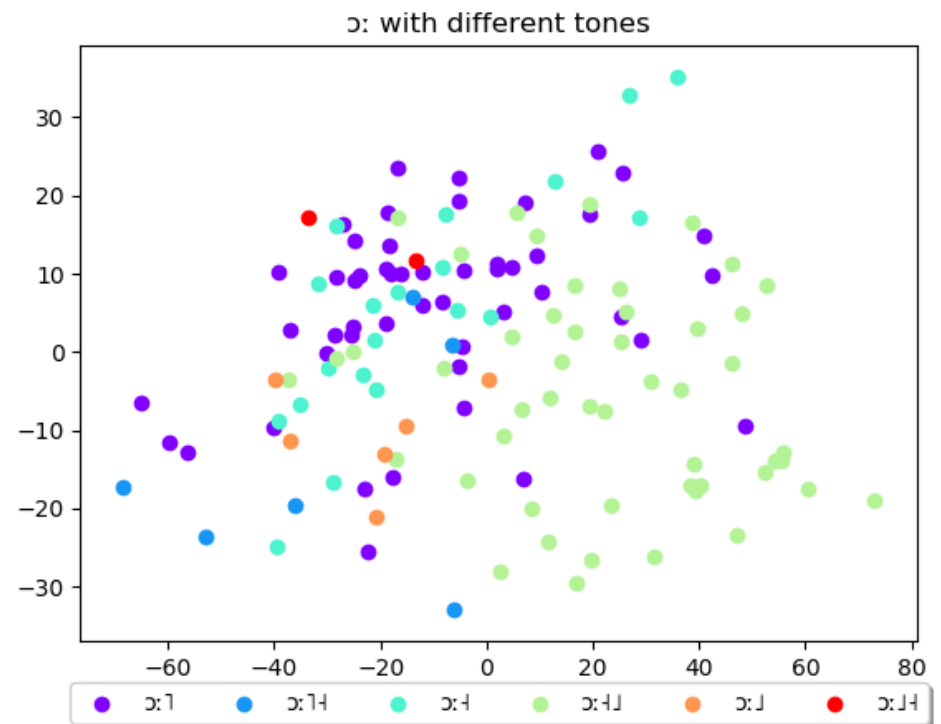
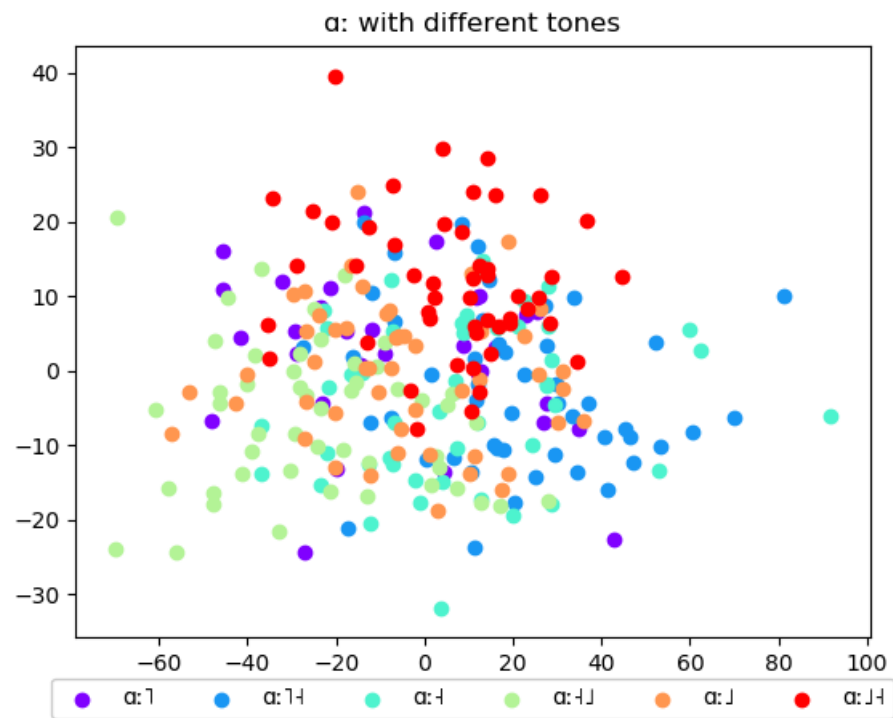
# Discussions

- Tone tier



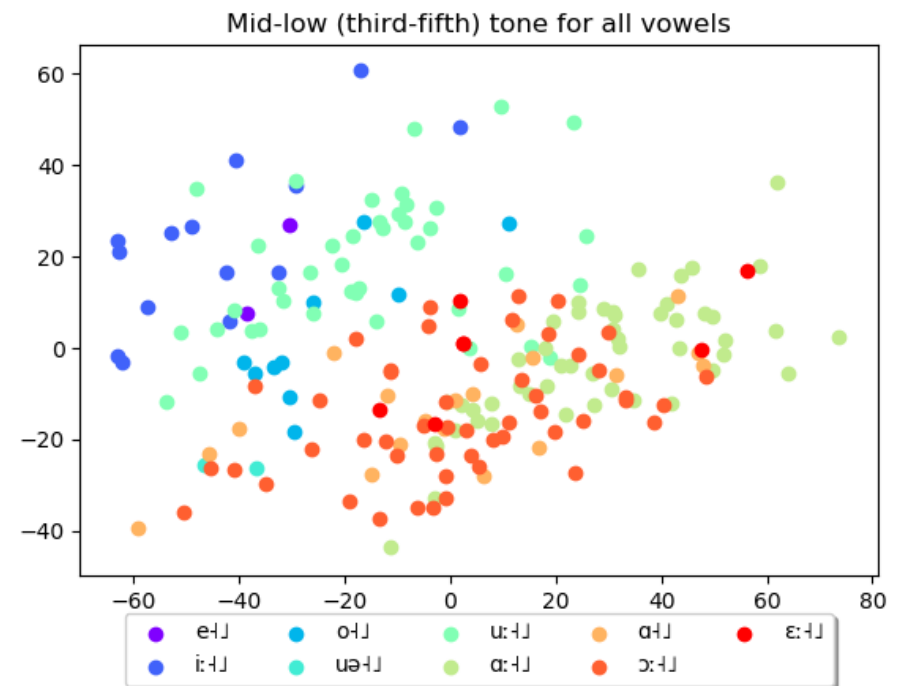
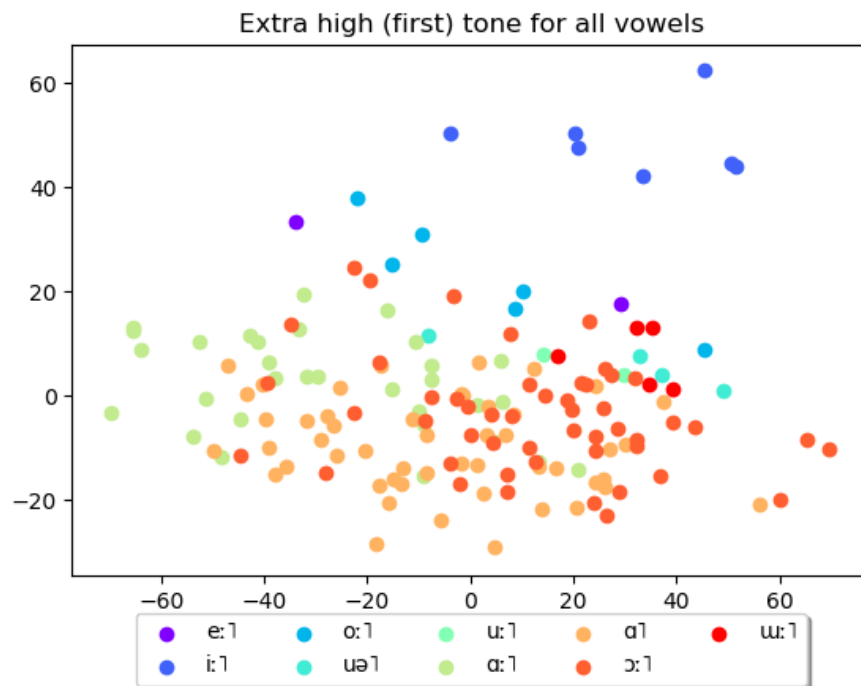
# Discussions

- Joint tier: single vowel with different tones



# Discussions

- Joint tier: single tone with different vowels



# Future work

- Improve phone error rates for consonants
- Methods of utilizing decoding sequences, such as combining vowels and consonants

# Thank you!